IT & C

ISSN 2821 - 8469, ISSN – L 2821 - 8469, Volumul 1, Numărul 1, Septembrie 2022

Megadatele (Big Data) pe Internet

Nicolae Sfetcu

Pentru a cita acest articol: Sfetcu, Nicolae (2022), Megadatele (Big Data) pe Internet, *IT & C*, **1**:1, 23-27, DOI: 10.58679/IT47091, https://www.internetmobile.ro/megadatele-big-data-pe-internet/

Publicat online: 19.08.2022

ABONARE

Megadatele (Big Data) pe Internet

Nicolae Sfetcu

Rezumat

Termenul Big Data se referă la extragerea, manipularea și analiza unor seturi de date care sunt prea mari pentru a fi tratate în mod obișnuit. Din această cauză se utilizează software special și, în multe cazuri, și calculatoare și echipamente hardware special dedicate. În general la aceste date analiza se face statistic. Pe baza analizei datelor respective se fac de obicei predicții ale unor grupuri de persoane sau alte entități, pe baza comportamentului acestora în diverse situații și folosind tehnici analitice avansate. Se pot identifica astfel tendințe, necesități și evoluții comportamentale ale acestor entități.

Cuvinte cheie: megadate, Big Data, Internet

IT & C, Volumul 1, Numărul 1, Septembrie 2022, pp. 23-27
ISSN 2821 - 8469, ISSN – L 2821 - 8469
URL: https://www.internetmobile.ro/megadatele-big-data-pe-internet/



Termenul Big Data se referă la extragerea, manipularea și analiza unor seturi de date care sunt prea mari pentru a fi tratate în mod obișnuit. Din această cauză se utilizează software special și, în multe cazuri, și calculatoare și echipamente hardware special dedicate. În general la aceste date analiza se face statistic. Pe baza analizei datelor respective se fac de obicei predicții ale unor grupuri de persoane sau alte entități, pe baza comportamentului acestora în diverse situații și folosind tehnici analitice avansate. Se pot identifica astfel tendințe, necesități și evoluții comportamentale ale acestor entități. Oamenii de știință folosesc aceste date pentru cercetări în meteorologie, genomică, (Nature 2008) conectomică, simulări fizice complexe, biologie, protecția mediului , etc. (Reichman, Jones, and Schildhauer 2011)

Odată cu creșterea volumului de date pe Internet, în media socială, cloud computing, dispozitive mobile și date guvernamentale, Big Data devine în același timp o amenințare și o oportunitate pentru cercetători în ceea ce privește gestionarea și utilizarea acestor date, menținând în același timp drepturile persoanelor implicate.

Definiții

Big Data includ, de obicei, seturi de date cu dimensiuni care depășesc capacitatea software și hardware obișnuite, folosind date nestructurate, semi-structurate și structurate, cu accentul pe datele nestructurate. (Dedić and Stanier 2017) Dimensiunile Big Data au crescut în timp din 2012, de la câteva zeci de terabyte până la multe exabyte de date. (Everts 2016) Eficientizarea lucrului cu Big Data implică învățarea mașinilor pentru a detecta modele, (Mayer-Schönberger and Cukier 2014) dar adesea aceste date sunt un produs secundar al altor activități digitale.

O definiție din 2018 afirmă că "**Big Data** sunt datele care necesită instrumentele de calcul paralel pentru a gestiona datele", aceasta reprezentând o turnură în informatică, prin utilizarea teoriilor de programare paralelă și lipsa unor garanții presupuse de modelele anterioare." Big Data utilizează statistici inductive și concepte de identificare a sistemelor neliniare pentru a deduce legi (regresii, relații neliniare și efecte cauzale) din seturi mari de date cu densitate scăzută de informații pentru a obține relații și dependențe sau pentru a efectua predicții ale rezultatelor și comportamentelor.

La nivelul Uniunii Europene nu există o definiție obligatorie dar, în conformitate cu Avizul 3/2013 al Grupului european de lucru privind protecția datelor,

> "**Big Data** este un termen care se referă la creșterea enormă a accesului și a utilizării automate a informațiilor: se referă la cantitățile uriașe de date digitale controlate de companii, autorități și alte organizații mari, care sunt supuse unor analize ample bazate pe utilizarea de algoritmi. Big Data pot fi folosite pentru a identifica tendințele și corelațiile generale, dar pot fi utilizate și pentru a afecta direct persoanele." (European Economic and Social Committee 2017)

Problema cu această definiție e că nu ia în considerare reutilizarea datelor cu caracter personal.

Regulamentul nr. 2016/679 definește **datele personale** (articolul 4, paragraful 1) drept

> "orice informație referitoare la o persoană fizică identificată sau identificabilă (persoana vizată); o persoană fizică identificabilă este cea care poate fi identificată, în mod direct sau indirect, în special prin referire la un identificator cum ar fi un nume, un număr de identificare, date de localizare, un identificator online sau unul sau mai mulți factori specifici identității fizice, fiziologice, genetice, mentale, economice, culturale sau sociale a acelei persoane fizice."

Definiția se aplică, la nivelul UE, și persoanelor neidentificate dar care pot fi identificate prin corelarea datelor anonime cu alte informații suplimentare. Datele cu caracter personal, o dată

anonimizate (sau pseudo-anonimizate), pot fi prelucrate fără a fi nevoie de o autorizație, ținându-se totuși cont de riscul re-identificării persoanei vizate.

Dimensiunile Big Data

Datele sunt partajate și stocate pe servere, prin interacțiunea dintre entitatea implicată și sistemul de stocare. În acest context, Big Data se poate clasifica în sisteme active (interacțiune sincronă, datele entității sunt trimise direct către sistemul de stocare), și sisteme pasive (interacțiune asincronă, datele sunt colectate printr-un intermediar și apoi introduse în sistem.

De asemenea, datele pot fi transmise direct în mod conștient, sau ne-conștient (dacă persoana ale cărei date sunt transmise nu este notificată la timp și clar). Datele sunt apoi prelucrate pentru a genera statistici.

În funcție de ținta analizelor statisticilor respective, dimensiunile datelor pot fi a) individuale (este analizat o singur entitate); sociale (se analizează grupuri discrete de entități din cadrul unei populații; și hibride (când o entitate este analizată prin prisma apartenenței sale la un grup deja definit).

Producția actuală imensă de date generate de utilizatori este estimată că va crește cu 2000% 1 2 3 la nivel mondial până în 2020, și sunt adesea nestructurate. (a7) În general, Big Data se caracterizează prin:

- Volum (cantitatea de date);
- Varietate (produse de diferite surse în diferite formate);
- Viteză (viteza de analiza online a datelor);
- Veracitate (datele sunt incerte și trebuie verificate);
- Valoare (evaluată prin analiză).

Volumul de date produse și stocate evoluează în prezent exponențial, peste 90% din ele fiind generate în ultimii patru ani. (European Economic and Social Committee 2017) Volumele mari necesită viteză mare de analiză, cu impact puternic asupra veracității. Datele incorecte au potențialul de a genera probleme atunci când sunt folosite în procesul de decizie.

Una din probleme important cu Big Data este dacă este nevoie de datele complete pentru a trage anumite concluzii cu privire la proprietățile lor, sau este suficient un eșantion. Big Data

conține chiar în nume un termen legat de dimensiune, care este o caracteristică importantă a Big Data. Dar eșantionarea (statistică) permite selectarea unor puncte corecte de colectare de date dintr-un set mai larg pentru a estima caracteristicile întregii populații. Big Data pot fi eșantionate pe diferite categorii de date în procesul de selecție a probelor cu ajutorul unor algoritmii de eșantionare pentru Big Data.

Bibliografie

Dedić, Nedim, and Clare Stanier. 2017. "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery." In *Innovations in Enterprise Information Systems Management and Engineering*, edited by Felix Piazolo, Verena Geist, Lars Brehm, and Rainer Schmidt, 114–22. Lecture Notes in Business Information Processing. Springer International Publishing.

European Economic and Social Committee. 2017. "The Ethics of Big Data: Balancing Economic Benefits and Ethical Questions of Big Data in the EU Policy Context." European Economic and Social Committee. February 22, 2017. https://www.eesc.europa.eu/en/our-work/publications-other-work/publications/ethics-big-data.

Everts, Sarah. 2016. "Information Overload." Science History Institute. July 18, 2016. https://www.sciencehistory.org/distillations/magazine/information-overload.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Reprint edition. Boston: Eamon Dolan/Mariner Books.

Nature. 2008. "Community Cleverness Required." *Nature* 455 (7209): 1. https://doi.org/10.1038/455001a.

Reichman, O. J., Matthew B. Jones, and Mark P. Schildhauer. 2011. "Challenges and Opportunities of Open Data in Ecology." *Science* 331 (February): 703. https://doi.org/10.1126/science.1197962.